Classification de fonctions continues à l'aide d'une distribution et d'une densité définies dans un espace de dimension infinie

Cuvelier, Etienne; Fraiture, Monique Noirhomme

Published in:
7èmes Jounées Francophones d'Extraction et de Gestion de Connaissances (EGC 07), Namur, 2007

Publication date:
2007

Document Version
Early version, also known as pre-print

Classification de fonctions continues à l'aide d'une distribution et d'une densité définies dans un espace de dimension infinie

Etienne Cuvelier*, Monique Noirhomme-Fraiture*

* Institut d'Informatique
Facultés Universitaires Notre-Dame de la Paix (FUNDP)
Namur, Belgique
{cuvelier.etienne,noirhomme.monique}@info.fundp.ac.be

Résumé. Il n'est pas rare que des données individu soient caractérisées par une distribution continue et non une seule valeur. Ces données fonctionnelles peuvent être utilisées pour classer les individus. Une solution élémentaire est de réduire les distributions à leurs moyennes et variances. Une solution plus riche a été proposée par Diday (2002) et mise en oeuvre par Vrac et al. (2001) et Cuvelier et Noirhomme-Fraiture (2005). Elle utilise des points de coupures dans les distributions et modélise ces valeurs conjointes par une distribution multidimensionnelle construite à l'aide d'une copule. Nous avons montré dans un précédent travail que, si cette technique apporte de bons résultats, la qualité de la classification dépend néanmoins du nombre et de l'emplacement des coupures. Les questions du choix du nombre et de l'emplacement des coupures restaient des questions ouvertes. Nous proposons une solution à ces questions, lorsque le nombre de coupures tend vers l'infini, en proposant une nouvelle distribution de probabilité adaptée à l'espace de dimension infinie que forment les données fonctionnelles. Nous proposons aussi une densité de probabilité adaptée à la nature de cette distribution en utilisant la dérivée directionnelle de Gâteaux. La direction choisie pour cette dérivée est celle de la dispersion des fonctions à classer. Les résultats sont encourageants et offrent des perspectives multiples dans tous les domaines où une distribution de données fonctionnelles est nécessaire.

1 Introduction

En analyse de données symbolique (voir Bock et Diday (2000)) une variable peut, entre autre être décrite par une distribution de probabilité continue. La classification en K groupes de ces données fonctionnelles peut être obtenue en utilisant une décomposition de mélange. Mais cette technique nécessite de pouvoir calculer la densité d'une distribution de fonction. Or l'espace des fonctions n'est pas un espace de dimension finie, tels que ceux où sont définies les distributions classiques. Projeter les fonctions dans un espace multidimensionnel par échantillonnage (voir Diday (2002)) permet de contourner ce problème, pour autant que l'on choisisse des distributions conjointes adéquates. Dans la section 2 de cet article nous rappel-

lerons brièvement la décomposition de mélange ainsi que l'algorithme des nuées dynamiques. Ensuite nous préciserons le cadre des distributions de fonctions et la construction de lois de ce type via les distributions multivariées. Nous terminerons ensuite, avant les conclusions, par une utilisation des nouveaux objets mathématiques définis pour la classification de données symboliques synthétiques.

2 Décomposition de mélange

2.1 Mélange de distributions

La décomposition de mélange est un outil important en classification. Elle consiste en l'estimation de la densité de probabilité qui est supposée avoir gouverné la génération d'un échantillon de données consitué de plusieurs groupes :

$$f(u) = \sum_{i=1}^{K} p_i \cdot f(u, \beta_i) \tag{1}$$

où les p_i représentent les proportions de chacun des groupes (leur somme étant égale à 1), et les fonctions $f(., \beta)$ les densités de ces groupes. Chaque composante du mélange correspondant en fait à un groupe. Pour trouver la partition $P = (P_1, ..., P_K)$ la mieux adaptée aux données deux grands algorithmes ont été proposés : EM (Estimation,Maximisation) par Dempster et al. (1977) et l'algorithme des nuées dynamiques par Diday et al. (1974). Nous avons choisi d'utiliser ce dernier car il avait déjà été utilisé dans le cadre de l'Analyse Symbolique par Diday (2002).

2.2 Algorithme des nuées dynamiques

L'algorithme utilisé est en fait une extension de la méthode des nuées dynamiques (Diday et al., 1974) dans le cas d'un mélange. L'idée principale est, alternativement, d'estimer au mieux la distribution de chaque classe, et ensuite de vérifier que chaque objet symbolique appartient à la classe de densité maximale. L'étape d'estimation est réalisée en maximisant un critère de qualité, ici la log-vraisemblance :

$$lvc(P, \beta) = \sum_{i}^{K} \sum_{u \in P_i} log(f(u, \beta_i)) \tag{2}$$

La classification commence avec une partition initiale aléatoire, et les deux étapes suivantes sont donc répétées jusqu'à stabilisation de la partition :

- **Etape 1 : Estimation des paramètres**
  Déterminer le vecteur $(\beta_1, ..., \beta_K)$ qui maximise le critère de qualité.
- **Etape 2 : Distribution des objets symboliques dans les classes**
  Les classes $(P_i)_{i=1,...,K}$, dont les paramètres ont été calculés à l'étape 1, sont construites comme suit
  $$P_i = \{u : f(u, \beta_i) \geq f(u, \beta_m) \forall m\}$$

Cet algorithme nécessite donc de pouvoir calculer la distribution, ou plus précisément la densité de probabilité, des objets à classer. Nous avons donc besoin de préciser la notion de distribution de fonctions.

3 Distribution de fonctions

3.1 Définitions

Définition 3.1 *Soit $\mathcal{D} = [a, b] \subseteq \mathbb{R}$ un intervalle fermé de $\mathbb{R}$, et $C^0(\mathcal{D})$ l'ensemble des fonctions continues bornées de domaine $\mathcal{D}$. Soient $u, v \in C^0(\mathcal{D})$, on définit :*

- $\|u\|_p = \left\{ \int_{\mathcal{D}} |u(x)|^p \, dx \right\}^{1/p}$
- $d_p(u, v) = \|u - v\|_p$
- $L^p(\mathcal{D}) = \left\{ u \in C^0(\mathcal{D}) \ : \ \|u\|_p < \infty \right\}$

Définition 3.2 *Soit Ω l'ensemble des objets dont les propriétés peuvent être décrites par une fonction de $L^2(\mathcal{D})$. Une* variable aléatoire fonctionelle (vaf) *$\underline{X}$ est définie comme étant toute fonction telle que :*

$$\underline{X} : \Omega \to L^2(\mathcal{D}) : \omega \mapsto X(\omega) \tag{3}$$

$$X(\omega) : \mathcal{D} \to \mathbb{R} : r \mapsto X(\omega)(r) \tag{4}$$

Définition 3.3 *Soient $f, g \in L^2(\mathcal{D})$. L'ordre ponctuel entre f et g sur l'intervalle $\mathcal{D}$ est défini par :*

$$\forall x \in \mathcal{D}, f(x) \leq g(x) \iff f \leq_{\mathcal{D}} g \tag{5}$$

Définition 3.4 *La* fonction de répartition fonctionnelle *ou* distribution fonctionnelle *d'une* vaf *$\underline{X}$ pour l'intervalle $\mathcal{D}$ est la fonction définie sur $L^2(\mathcal{D})$ par :*

$$\begin{aligned} F_{\underline{X},\mathcal{D}}(u) &= P\{\omega \in \Omega : X(\omega) \leq_{\mathcal{D}} u\} \\ &= P[\underline{X} \leq_{\mathcal{D}} u] \end{aligned} \tag{6}$$

où $u \in L^2(\mathcal{D})$.

Si la notion de distribution de fonction est facile à définir, il paraît, par contre, plus malaisé de donner immédiatement un moyen de la calculer. Considérons l'exemple de la figure 1. Supposons que les lignes continues forment un échantillon fonctionnel homogène. Si v est une de ces fonctions, calculer $F_{\underline{X},\mathcal{D}}(v)$ peut se faire empiriquement :

$$\widehat{F}_{\underline{X},\mathcal{D}}(v) = \frac{\#\{f \in A : f \leq_{\mathcal{D}} v\}}{\#A}$$

Mais qu'en est-il pour les fonctions w et u ? Pour w on peut supposer intuitivement que la valeur de $F_{\underline{X},\mathcal{D}}(w)$ est proche de 90%. Et pour u, est-ce 50%, car u est toujours supérieure à 10 des 20 fonctions de l'échantillon ? Et ce malgré le fait que u soit supérieure à 12 des 20 fonctions sur plus de la moitié du domaine ?
Pour solutionner ce problème de calcul, nous allons projeter dans un espace multidimensionnel les fonctions, par nature définies dans un espace de dimension infinie.

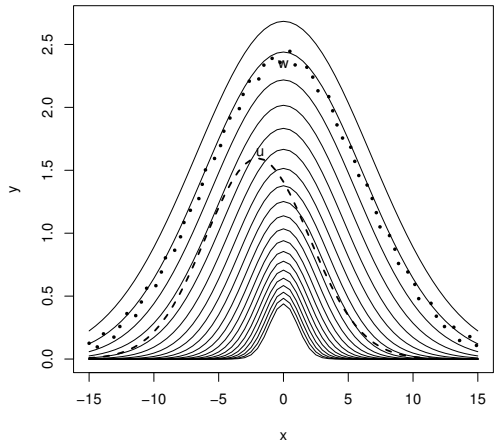

FIG. 1 – *Un exemple d'échantillon de 20 fonctions*

3.2 Approximations multivariées

3.2.1 Introduction

Soient $n \in \mathbb{N}$ et $q = 2^n + 1$, nous définissons :$\left\{x_1^n, \ldots, x_q^n\right\}$, q points équidistants de $\mathcal{D}$, avec $x_1^n = a$ and $x_q^n = b$. Bien sûr nous avons que

$$\left|x_{i+1}^n - x_i^n\right| = \frac{|\mathcal{D}|}{2^n} = \frac{|\mathcal{D}|}{q} \ , \ \forall i \in \{1, \ldots, q-1\} \tag{7}$$

Si nous définissons ensuite les deux ensembles suivants :

$$\mathcal{A}(u) = \{\omega \in \Omega : X(\omega) \leq_{\mathcal{D}} u\}$$

$$\mathcal{A}_n(u) = \bigcap_{i=1}^{q} \{\omega \in \Omega : X(\omega)(x_i^n) \leq u(x_i^n)\}$$

alors nous pouvons utiliser l'approximation suivante :

$$F_{\underline{X},\mathcal{D}}(u) = P[\mathcal{A}(u)] \approx P[\mathcal{A}_n(u)] = H\left(u(x_1^n), \ldots, u(x_q^n)\right) \tag{8}$$

où H est une distribution multivariée de dimension q. Nous pouvons donc utiliser une distribution conjointe pour approximer notre *distribution fonctionnelle*. Le choix de la distribution, ou de la famille de distributions, à utiliser est évidemment important. Avant de préciser ce choix, remarquons que pour une valeur choisie $x \in \mathcal{D}$, il est très facile d'estimer la distribution des valeurs de $\underline{X}(x)$.

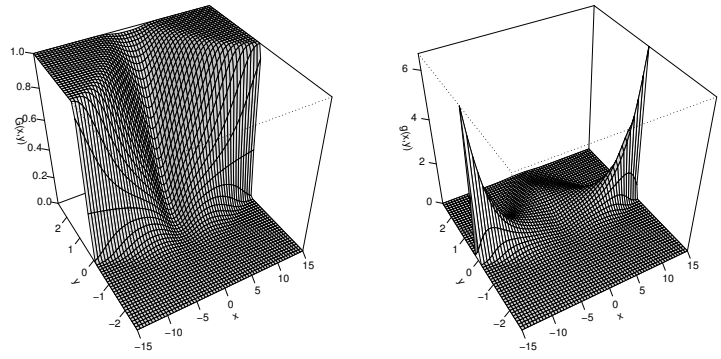

FIG. 2 – *Les surfaces $G(x, y)$ et $g(x, y)$ de l'exemple de la Fig. 1*

Définition 3.5 *Soit $\underline{X}$ une* vaf. *Les fonctions g et G, respectivement appelées* surface de distributions *et* surface de densités*, de domaines $\mathcal{D}$ et d'images $[0, 1]$ sont définies par*

$$G(x, y) = P[\underline{X}(x) \leq y] \qquad g(x, y) = \frac{\partial}{\partial x} G(x, y) \tag{9}$$

Il est assez facile de calculer G et g à l'aide des techniques univariées. Ainsi, si $\underline{X}$ est un processus Gaussien, alors ces deux fonctions peuvent être calculées pour une valeur donnée de x par la fonction de répartition et la densité de la loi $\mathcal{N}(\mu(x), \sigma(x))$.
Dans les cas où l'on ignore la loi suivie par $\underline{X}(x)$ on utilisera l'estimation empirique pour G et l'estimation à noyaux pour g :

$$\hat{G}(x, y) = \frac{\#\{X_i(x) \leq y\}}{N} \qquad \hat{g}(x, y) = \frac{1}{N \cdot h(x)} \sum_{i=1}^{N} K\left(\frac{y - X_i(x)}{h(x)}\right) \tag{10}$$

La Fig. 2 montre ces deux surfaces avec l'exemple de la Fig. 1, dans le cas Gaussien. Etant donné qu'il est très facile de calculer les marges de la distribution H par :

$$G(x_1^n, u(x_1^n)), \ldots, G\left(x_q^n, u\left(x_q^n\right)\right)$$

l'idée de reconstruire cette distribution H à partir de ses marges a été proposée par Diday (2002) en utilisant les copules archimédiennes.

3.2.2 Copules archimédiennes

Définition 3.6 *Une copule C est une distribution multivariée définie sur le cube $[0, 1]^n$, dont toutes les marginales sont uniformes sur $[0, 1]$.*

$$C : [0, 1]^n \to [0, 1] : (u_1, \ldots, u_n) \mapsto C(u_1, \ldots, u_n)$$

Les copules sont des outils précieux dans la modélisation des structures de dépendance grâce au théorème de Sklar (voir Nelsen (1999)).

TAB. 1 – *Générateurs archimédiens*

Nom	**Générateur**	**Dom.** θ
Clayton	$t^{\theta}-1$	$\theta > 0$
Frank	$-\ln \frac{e^{-\theta \cdot t}-1}{e^{-\theta}-1}$	$\theta > 0$
Gumbel-Hougaard	$(-\ln t)^{\theta}$	$\theta \geq 1$

Théoreme 3.1 (Sklar) *Si $H(x_1, \ldots, x_n)$ est une distribution multivariée de marges $F_1(x_1)$,...,$F_n(x_n)$, alors il existe une copule C telle que*

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)) \tag{11}$$

De plus, si $F_1, ..., F_n$ sont toutes continues, alors C est unique ; sinon C est unique seulement sur $domF_1 \times ... \times domF_n$.

Définition 3.7 *Les copules archimédiennes sont définies par*

$$C(\underline{u}) = C(u_1, ..., u_n) = \psi\left(\sum_{i=1}^{n} \phi(u_i)\right) \tag{12}$$

où ϕ appelé générateur, est une fonction continue strictement décroissante de $[0,1]$ vers $[0,\infty]$ telle que :

- *$\phi(0) = \infty$ et $\phi(1) = 0$*
- *$\psi = \phi^{-1}$ soit complètement monotonique sur $[0,\infty[$ c-à-d que $\forall t \in [0,\infty[$ et $\forall k \geq 0$*

$$(-1)^k \psi^{[k]}(t) \geq 0$$

où $\psi^{[k]}$ représente la dérivée d'odre k de ψ.

Le tableau 1 montre trois familles de générateurs Archimédiens. Si nous utilisons conjointement les *surfaces de distributions* et les copules archimédiennes, alors notre approximation (8) peut directement se récrire :

$$F_{\underline{X},\mathcal{D}}(u) \approx P[\mathcal{A}_n(u)] = \psi\left(\sum_{i=1}^{q} \phi\left(G\left[x_i^n, u\left(x_i^n\right)\right]\right)\right) \tag{13}$$

La densité conjointe étant donnée par l'expression suivante :

$$\frac{\partial^q}{\partial u_1 ... \partial u_q} C(G\left((x_1^n), u(x_1^n)\right), \ldots, G\left((x_q^n), u(x_q^n)\right)) \cdot \prod_{i=1}^{q} g\left((x_i^n), u(x_i^n)\right) \tag{14}$$

L'approximation (13) a déjà été utilisée en deux dimensions avec la copule de Frank par Vrac et al. (2001), et pour un nombre quelconque de dimensions dans avec la copule de Clayton Cuvelier et Noirhomme-Fraiture (2005). A la suite de ces travaux deux questions restaient sans réponse :

1. Quelle valeur de q choisir ?

2. Où choisir les valeurs $x_1^n, \ldots, x_q^n$?

Le choix de points équidistants semble répondre à la première question même si nous verrons qu'une précision doit encore être apportée. Pour la seconde question, sur une idée d'Edwin Diday, nous nous proposons de considérer l'évolution naturelle de cette approximation : la limite lorsque q tend vers l'infini.
Avant d'aller plus loin remarquons que, pour une probabilité $p \in [0, 1[$, on a que $\forall 0 \leq \epsilon \leq 1$:

$$q \geq \frac{\phi(\epsilon)}{\phi(p)} \Rightarrow \psi \left[\sum_{i=1}^{q} \phi(p) \right] \leq \epsilon$$

Cela signifie que la limite de l'expression (13), lorsque $q \to \infty$ est presque toujours nulle ! Pour éviter ce problème, nous proposons d'utiliser un nouveau type de distributions basée sur les moyennes quasi-arithmétiques.

3.2.3 Moyennes quasi-arithmétiques discrètes

Définition 3.8 *Soient $[a, b]$ un intervalle réel, et $q \in \mathbb{N}_0$. Une moyenne quasi-arithmetique est toute fonction $M : [a, b]^q \to [a, b]$ telle que :*

$$M(x_1, \ldots, x_q) = \psi \left(\frac{1}{q} \sum_{i=1}^{q} \phi(x_i) \right) \tag{15}$$

où ϕ est une fonction continue strictement monotone, et $\psi = \phi^{-1}$.

Le concept de moyenne quasi-arithmétique a été introduit par Kolmogorov (1930) et Nagumo (1930), et a été étudié dans le cadre des équations fonctionnelles par Aczel (1966).

Proposition 3.2 *Soient $q \in \mathbb{N}_0$, $\{F_i | 1 \leq i \leq q\}$ un ensemble de distributions univariées et ϕ générateur Archimédien, alors*

$$H(x_1, \ldots, x_q) = \psi \left(\frac{1}{q} \sum_{i=1}^{q} \phi(F_i(x_i)) \right) \tag{16}$$

est une distribution conjointe de marges

$$F_i^*(x) = \psi \left(\frac{1}{q} \cdot \phi(F_i(x)) \right) \tag{17}$$

Nous appelons cette distribution Moyenne Quasi-Arithmétique de Marges *(en anglais :*Quasi-Arithmetic Mean of Margins (QAMM)*).*

Démonstration Il suffit de remarquer que si F_i est une distribution univariée, alors F_i^* aussi, et d'ensuite utiliser ces nouvelles distributions et la copule générée par ϕ pour construire la distribution multivariée.

3.3 Distribution et densité définies dans un espace de dimension infinie

3.3.1 Moyennes quasi-arithmétiques continues

En utilisant l'expression (16) et en notant $\left|x_{i+1}^n - x_i^n\right| = \Delta_x$ (cf. (7)) $\forall i$ on peut écrire :

$$\begin{aligned}
\lim_{n\to\infty} P[\mathcal{A}_n(u)] &= \lim_{n\to\infty} \psi\left[\frac{1}{q}\sum_{i=1}^{q} \phi\left(G\left[x_i^n, u(x_i^n)\right]\right)\right] \\
&= \lim_{n\to\infty} \psi\left[\frac{1}{|\mathcal{D}|}\sum_{i=1}^{q} \frac{|\mathcal{D}|}{q} \cdot \phi\left(G\left[x_i^n, u(x_i^n)\right]\right)\right] \\
&= \psi\left[\frac{1}{|\mathcal{D}|} \cdot \int_{\mathcal{D}} \phi\left(G\left[x, u\left(x\right)\right]\right) dx\right] \qquad (18)
\end{aligned}$$

Définition 3.9 *Soient $\underline{X}$ une* vaf*, $u \in L^2(\mathcal{D})$, G sa* Surface de Distributions *et ϕ un générateur archimédien. Nous appellerons l'expression* (18) Moyenne Quasi-Arithmétique Continue de Marges *(en anglais* Quasi-Arithmetic Mean of Margins Limit (QAMML)*).*

De Finetti (1931) et Hardy et al. (1934) sont les premiers à avoir étendu au cas continu les résultats de Nagumo et Kolmogorov. De par les propriétés des moyennes la limite (18) existe toujours et est comprise entre 0 et 1. Cette distribution définie sur espace de dimension, par nature, infinie, ne sera véritablement utile que si elle est munie d'une densité. Il n'est évidemment plus possible d'utiliser l'expression (14) comme dans le cas fini. Nous proposons donc d'utiliser une densité "directionnelle".

3.3.2 Densité de Gâteaux

Rappelons ici un concept provenant de l'analyse fonctionnelle : la dérivée de Gâteaux, qui est une dérivée directionnelle (cf. Atkinson et Han (2001)).

Définition 3.10 *Soient V et W deux espaces vectoriels normés, et F un opérateur de V vers W. La* differentielle de Gâteaux *$DF(u;s)$ de F en u dans la direction $s \in V$ est donnée par :*

$$\begin{aligned}
DF(u;s) &= \lim_{\epsilon\to 0} \frac{F(u+\epsilon\cdot s) - F(u)}{\epsilon} \qquad (19) \\
&= F'(u)\cdot s \qquad (20)
\end{aligned}$$

Si la limite (19) *existe $\forall s \in L^2(\mathcal{D})$ alors F est dite* Gâteaux différentiable *et l'application $F'(u)$ est la* dérivée de Gâteaux *de F en u.*

L'utilisation de ce type de différentiation nécessite donc de préciser dans quelle direction elle se fait. Nous proposons d'utiliser comme fonction de direction toute fonction permettant de mesurer la dispersion des données pour toute valeur de x, avec comme exemple le plus immédiat l'écart-type σ.

Définition 3.11 *Soient $\underline{X}$ une* vaf*, $F_{\underline{X},\mathcal{D}}$ sa distribution fonctionnelle et u une fonction de $L^2(\mathcal{D})$. Si $s \in L^2(\mathcal{D})$ est une fonction telle que $s(x)$ mesure la dispersion des valeurs de*

$\underline{X}(x)$*, alors nous définissons la* densité de Gâteaux de $F_{\underline{X},\mathcal{D}}$ en u et dans la direction s *par :*

$$\begin{aligned} f_{\underline{X},\mathcal{D},s}(u) &= \lim_{\epsilon \to 0} \frac{F_{\underline{X},\mathcal{D}}\left(u + s \cdot \epsilon\right) - F_{\underline{X},\mathcal{D}}\left(u\right)}{d_2\left(u + s.\epsilon, u\right)} \\ &= \frac{DF_{\underline{X},\mathcal{D}}(u;s)}{\|s\|_2} \end{aligned} \tag{21}$$

où $DF_{\underline{X},\mathcal{D}}(u;s)$ *est la* différentielle de Gâteaux *de* $F_{\underline{X},\mathcal{D}}$ *en* u *dans la direction* $s \in V$.

La dérivée de Gâteaux d'une transformée intégrale étant un résultat classique d'analyse fonctionnelle (cf. Lusternik et Sobolev (1974)), nous avons le résultat suivant.

Théoreme 3.3 *Soient* $F_{\underline{X},\mathcal{D}}$ *, la Moyenne Quasi-Arithmétique Continue de Marges d'une fonction* u *de* $L^2(\mathcal{D})$*. Si* $s \in L^2(\mathcal{D})$ *est une mesure fonctionnelle de la dispersion des valeurs de* $\underline{X}(x)$*, alors la* densité de Gâteaux de $F_{\underline{X},\mathcal{D}}$ calculée en u dans la direction de s *est donnée par :*

$$\begin{aligned} f_{\underline{X},\mathcal{D},s}(u) &= \frac{1}{\|s\|_2 \cdot |\mathcal{D}|} \psi' \left[\frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \phi\left(G\left[t, u(t)\right]\right) \, dt \right] \\ & \quad \left\{ \int_{\mathcal{D}} \phi'\left(G\left[t, u(t)\right]\right) g\left[t, u(t)\right] s(t) \, dt \right\} \end{aligned} \tag{22}$$

Soulignons ici l'intérêt de diviser la différentielle de Gâteaux dans l'expression (21) par la norme de s. En effet, comme on peut le constater dans (22), sans cela, la densité de Gâteaux d'une Moyenne Quasi-Arithmétique Continue de Marges calculée avec deux paramétrages différents $s_1 \leq_{\mathcal{D}} s_2$ pourrait donner la même valeur, pour autant que $g_1\left[t, u(t)\right] s_1(t) = g_2\left[t, u(t)\right] s_2(t)$ pour toute valeur de $t \in \mathcal{D}$. La division par la norme de la mesure de dispersion permet de réintroduire cette distinction.

3.4 Domaines des modèles

Remarquons maintenant que le calcul de l'expression (22) nécessite de pouvoir calculer $g\left(x, y\right)$ sur l'ensemble des valeurs de $\mathcal{D}$ et que, ceci n'est en général possible que si la mesure de dispersion s est non nulle. Nous dirons que le domaine du modèle est l'ensemble des réels pour lesquels $s(x) > 0$. Nous appellerons donc *domaine du modèle* tout intervalle $\mathcal{D} \subseteq \{x \in \mathbb{R} : s(x) > 0\}$.
Ainsi, dans le cas des Moyennes Quasi-Arithmétiques de Marges utilisées conjointement avec la densité de Gâteaux, nous répondons aux deux questions évoquées plus avant concernant le nombre et les choix des points $x_1^n, \ldots, x_q^n$:

1. quand à la valeur de q : on le choisit très grand (QAMM), voire on le fait tendre vers l'infini (QAMML) pour minimiser l'erreur due à l'approximation,
2. quand au choix des $x_1^n, \ldots, x_q^n$: ils doivent se situer dans le domaine du modèle, c'est-à-dire pour les valeurs de x où il y a dispersion non nulle des valeurs de $\underline{X}(x)$.

Il faut noter que ces deux règles peuvent aussi s'appliquer dans le cas d'utilisation des copules et de la densité multivariée. En effet le même problème de calculabilité se pose avec l'expression (14) si la dispersion des valeurs $\underline{X}(x)$ est nulle pour au moins une des dimensions. Mais

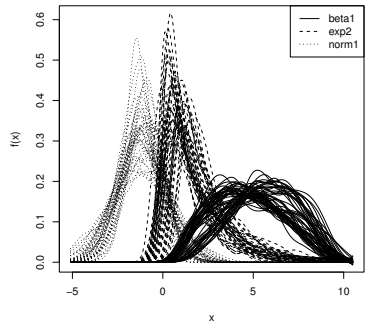

FIG. 3 – *Les 140 données fonctionnelles du test*

il est évidemment conceptuellement plus difficile de définir la notion de domaine de modèle dans le cas multivarié, car cela équivaut à ne pas toujours utiliser le même nombre de dimensions (et pas nécessairement les mêmes) pour calculer une même distribution ou sa densité en plusieurs endroits. Sauf, si l'on se souvient que nous ne sommes pas confrontés à de vraies données multivariées, mais à la projection en dimension q de données définies dans un espace de dimension infinie.

4 Application

Nous avons donc utilisé les Moyennes Quasi-Arithmétiques Continues de Marges, conjointement avec la densité de Gâteaux, dans le cadre de l'algorithme des nuées dynamiques sur des données de type symbolique : des densités de probabilités. Pour notre test nous avons utilisé un ensemble de 140 données synthétiques mixant des exponentielles, des normales et des bétas (Fig. 3). Pour constituer cet ensemble de données, pour chaque distribution nous avons généré 500 nombres aléatoires suivant la loi choisie et ensuite nous avons réalisé une estimation à noyaux à partir de ces nombres. Les résultat fonctionnel étant stocké à l'aide des fonctions splines.

Remarquons que les distributions de probabilités sont des données fonctionnelles qui sont définies sur $\mathbb{R}$, même si la valeur de la fonction peut être nulle sur une partie du domaine (exemple :la loi exponentielle). Or, en classification, seule une partie du domaine de la fonction est intéressante : celle où l'on peut distinguer cette fonction des autres, c'est-à-dire là où la fonction est non nulle (ou supérieure à ϵ fixé). Nous restreignons donc, pour des raison classificatoires, le domaine sur lequel nous calculons la Moyenne Quasi-Arithmétique Continue de Marges (ou sa densité de Gâteaux) aux valeurs distinguables de la fonction considérée. Pour cela nous utilisons une fonction de "confiance" : $\tau : \cup_{i=1}^{K} \mathcal{D}_i \rightarrow \{0, 1\}$:

$$\begin{aligned} \tau(x) &= 1 \ si \ x > 0 \\ &= 0 \ sinon \end{aligned}$$

Avec cette fonction l'expression (18) devient :

$$\psi\left[\frac{1}{\int_{\mathcal{D}}\tau(t)dt}\int_{\mathcal{D}}\phi\left(G\left[x,u\left(x\right)\right]\right)\cdot\tau(u(x))dx\right] \tag{23}$$

L'implémentation de l'algorithme et des lois *QAMML* ont été réalisées à l'aide du logiciel R (R Development Core Team (2005)). Nous avons utilisé le générateur de *Clayton* (cf. Table 1), les estimations $\hat{G}$ et $\hat{g}$ (expressions (10)) et l'écart-type σ comme direction pour les densités de Gâteaux. Nous avons exécuté la méthode cinq fois et nous avons retenu le résultat avec la meilleure valeur du critère, et nous avons obtenu un taux de mauvaise classification de 7.1%, c-à-d 10 données fonctionnelles mal classées sur 140.

5 Conclusions

Dans cet article nous proposons d'utiliser deux outils mathématiques nouveaux, les Moyennes Quasi-Arithmétiques de Marges et la densité de Gâteaux, dans le cadre de la décomposition de mélange classifiante. Ces outils nous permettent d'apporter une réponse aux questions laissées sans réponse par les travaux précédents : à savoir le nombre et le choix des points d'approximation. L'utilisation de ces nouveaux objets mathématiques dans le cadre de la classification non supervisée sur des données symboliques synthétiques donne des résultats encourageants. D'autres outils et méthodes de classification de données fonctionnelles existent, mais l'utilisation de distributions adéquates permet d'obtenir une modélisation probabiliste des données. D'autre part un certain nombre de développements sont encore envisageables pour affiner cet outil mathématique : l'utilisation dans (18) d'une distribution autre que la distribution uniforme sur $\mathcal{D}$, l'utilisation conjointe de la distribution des dérivées successives d'une fonction u ou encore l'utilisation de directions autres que s, plus discriminantes (cf. Ramsay et Siverman (2005)).

Références

Aczel, J. (1966). *Lectures on Functional Equations and Their Applications*. Mathematics in Science and Engineering. New York and London : Academic Press.

Atkinson, K. et W. Han (2001). *Theoretical Numerical Analysis*. texts in Applied Mathematics. New-York : Springer.

Bock, H. et E. Diday (2000). *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer Verlag.

Cuvelier, E. et M. Noirhomme-Fraiture (2005). Clayton copula and mixture decomposition. In *ASMDA 2005*, pp. 699–708.

De Finetti, B. (1931). Sul concetto di media. *Giornale dell' Instituto Itialiano degli Attuari 2*, 369–396.

Dempster, A. P., N. M. Laird, et D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society (Series B) 39*(1), 1–38.

Diday, E. (2002). Mixture decomposition of distributions by copulas. In *Classification, Clustering and Data Analysis*, pp. 297–310.

Diday, E., A. Schroeder, et Y. Ok (1974). The dynamic clusters method in pattern recognition. In *IFIP Congress*, pp. 691–697.

Hardy, G. H., J. E. Littlewood, et G. Polya (1934). *Inequalities*. Cambridge: Cambridge University Press.

Kolmogorov, A. (1930). Sur la notion de moyenne. *Rendiconti Accademia dei Lincei 12*(6), 388–391.

Lusternik, L. A. et V. J. Sobolev (1974). *Elements of Functional Analysis*. Delhi: Hindustan Publishing Corpn.

Nagumo, M. (1930). Uber eine klasse der mittelwerte. *Japan Journal of Mathematics 7*, 71–79.

Nelsen, R. (1999). *An introduction to copulas*. London: Springer.

R Development Core Team (2005). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ramsay, J. O. et B. W. Siverman (2005). *Functionnal Data Analysis*. Springer Series in Statistics. New-York: Springer.

Vrac, M., E. Diday, A. Chédin, et P. Naveau (2001). Mélange de distributions de distributions, décomposition de mélange de copules et application à la climatologie. In *Actes du VIIIème congrés de la Société Francophone de Classification*, pp. 348–355.

Summary

Individual data can be caracterized by continuous distributions and not by a single value. Those functional data can be used to classify individuals. In a elementary solution, we can reduce distribution to mean and variance. Another richer solution is proposed by Diday (2002) and implemented by Vrac et al. (2001) and Cuvelier et Noirhomme-Fraiture (2005). It uses cut points in the distributions and models those joint values by a multidimensional distribution built with copulas. We have shown in a previous work that even if this approach gives good results, classification quality depends on the number and the place of cutpoints. The questions of number and place of cuts remains open questions. We propose a solution to these questions, when the number of cuts tends to infinity. We suggest a new distribution adapted to the space of infinite dimension. We suggest also a density which uses the Gâteaux directional derivative. The chosen direction is dispersion of functions to be classified. Results are encouraging and offer multiples perspectives in all the domains where functional data distribution is necessary.